3/PRTS

METHOD FOR REDUCING CONGESTION IN A NETWORK

The field of the invention is that of communication networks. In order to communicate, data terminal equipment units (DTEs) use various protocols.

Normally, there are several communication layers, for example an application layer, a transport layer and a network layer. The application layer is not directly concerned with the locations where functions are executed. To give a non-limiting example, there are various protocols that are usable in an application layer, such as TELNET for coupling a local terminal with a remote machine, FTP for transferring files between machines, and HTTP for accessing web pages. Generally, a client application sends requests to a server application, from which it receives responses without being concerned with the fact that the server application may be running in a physical and logical environment different from that of the client application. The function of the transport layer is to allow two applications to communicate by accommodating the physical and logical environment of each of them. There are protocols for the transport layer that are connected, such as TCP, and protocols that are connectionless, such as UDP. The advantage of a connected transport protocol is that it ensures the reliability of the exchanges, while a connectionless protocol provides greater speed. The function of the network layer is to route the messages between two DTEs while adapting to the networks to which the two DTEs are connected. For example, network protocols such as IP or CLNP provide a connectionless service of the datagram type. This means that for a message composed of datagrams, the network protocol retransmits each datagram from machine to machine based on the availability of the paths offered, without ensuring that each datagram sent actually arrives at its destination, for example in case of congestion in a network or an intermediate machine. Such an intermediate machine, responsible for propagating the datagrams between two different networks, is generally called a router. According to the recommendation I.113 ITU-T "Vocabulary of Terms for Broadband Aspects of ISDN," Helsinki, March 1993, a connectionless service is defined as a service that allows the transfer of information between two users of the service without any need for procedures for establishing end-to-end calls.

When the number of datagrams to be propagated between two networks exceeds the transmission capacity of the router, the datagrams are placed in queues

inside the router in order to be processed later. When the number of datagrams waiting to be processed in a queue exceeds a threshold, the router discards any new datagrams that arrive that would take up space in this queue because the storage capacity of the router is limited.

The result of these delay and loss phenomena is that, in case of congestion, the transport layers of the DTEs re-send, increasing the number of datagrams to be propagated by the router and thus aggravating the congestion.

Known solutions exist at the transport layer level, such as for example the Slow Start with Congestion Avoidance of the TCP protocol. The transport layer of the sending DTE detects the network congestion when it realizes that it must retransmit the data. It relieves the routers that its traffic passes through by temporarily and spontaneously reducing its sending capacity for a given connection. This is done, for example, by increasing the time interval between two possible retransmissions, or by sending a smaller quantity of information than is acceptable by the receiving DTE. One drawback of this solution is that it is used when the congestion has already occurred, so retransmission is inevitable, and the need to delay it further considerably reduces communication performance.

Another example is the Source Quench of the ICMP control protocol. When the router realizes that the number of messages in a queue reaches an intermediate threshold the moment a new datagram is placed in the queue, it sends a particular message to the DTE that sent the datagram, in order to tell it that the risk of congestion has increased. The sending DTE then reduces its sending capacity. One drawback of this solution is that the sending DTE is not clearly informed as to when it can increase its speed again. Moreover, this solution makes it necessary to send additional messages through the networks.

In order to control the speed of the sending DTE so as to reduce it before congestion occurs, without generating any additional traffic, the invention uses datagrams that pass through the router in the receiving DTE-to-sending DTE direction, and that contain window information. This window information is generated at the transport protocol level by the receiving DTE in order to inform the sending DTE of the quantity of information that the latter is authorized to send through a connection prior to receiving an acknowledgement indicating that the preceding transmitted information has been correctly received.

The subject of the invention is a method for reducing the congestion in a network layer of a router machine when it accumulates in a queue datagrams to be sent through a network, characterized in that it comprises:

- a first step that measures a fullness level of said queue, in order to generate a signal based on said fullness level;

- a second step that detects any datagram received from said network, wherein a field of a transport layer contains a first received window value;

- a third step that generates a second sent window value based on said signal, in order to process the detected datagram by entering said second value into it in said field;

- a fourth step that routes the processed datagram through a network to a transport layer, which limits its send rate based on the sent window value.

Thus, in an environment with a network protocol that provides a connectionless datagram service and a transport protocol that provides a reliable connection service using a window system to control the flow from end to end, an intermediate machine through which all the datagrams exchanged by two data terminal equipment units pass can control the flow of datagrams passing through it by acting via its network layer on the transport layer of the sending data terminal equipment unit. This offers the advantage of reducing the congestion in the intermediate machine without requiring a particular procedure in the data terminal equipment units.

A preferred exemplary embodiment of the invention is explained in the following description in reference to the figures, in which:

- Fig. 1 shows a router of the prior art;
- Fig. 2 shows a transport layer segment;
- Fig. 3 shows a network layer datagram;
- Fig. 4 shows a router that implements the invention;
- Fig. 5 shows a datagram with window information;
- Fig. 6 shows some steps of the method according to the invention.

Referring to Fig. 1, a sending DTE (data terminal equipment) 1 communicates messages to a receiving DTE 2 using a transport protocol 3. To do this, a transport layer 4 of the DTE 1 generates segments 5 addressed to a transport layer 6 of the DTE 2. More generally, the segments are Transport Data Protocol Units TPDU.

Referring to Fig. 2, each segment 5 of the transport layer comprises at least one transport field 7. When the message relates to an application, the transport layer 4 receives information from an application layer (not represented) of the machine 1 via an interface 8. The transport layer 4 then incorporates this information into a field 9 of the segment 5.

Referring to Fig. 1, the transport layer 4 transmits the segment 5 to a network layer 10 of the DTE 1 via an interface 11.

Referring to Fig. 3, the network layer 10 juxtaposes a field 13 with the segment 5 so as to create a datagram 12 addressed to a network layer 14 of the receiving DTE 2. The field 13 contains data for implementing a network protocol between the DTEs 1 and 2. If, for example, the network is an IP network, the field 13 contains an IP address that identifies the receiving DTE 2 and an IP address that identifies the sending DTE 1.

Depending on the topological configuration of the network, the datagram 12 is routed directly from the DTE 1 to the DTE 2 or indirectly through one or more units of router equipment 15. Referring to Fig. 1, the network layer 10 routes the datagram 12 to a network layer 16 of a unit of router equipment 15, through a physical layer 17. The network layer 16 then routes the datagram 12 to the layer 14 through a physical layer 18. When the physical layer 18 is not directly connected to the DTE 2, the datagram 12 passes through as many units of router equipment as necessary in order to reach a physical layer to which the DTE 2 is directly connected.

When the network layer 16 of the router equipment 15 receives a datagram 19 that it cannot immediately retransmit through the physical layer 18 due to congestion, it accumulates the datagram 19 in a first queue 20, which it empties as the physical layer 18 becomes available.

When the network layer 14 of the DTE 2 receives the datagram 12, it extracts the field 13 from it in order to transfer it in the form of a segment 5 to the transport layer 6 via an interface 21. When the segment 5 comprises a field 9 addressed to an application, the transport layer 6 transfers the field 9 to an application layer (not represented) of the DTE 2 via an interface 22. In addition, the transport layer 6 sends the transport layer 4 an acknowledgement segment to inform it that it has actually received the segment 5. To do this, the transport layer 6 transmits the acknowledgement segment, which generally comprises only the field 7, to the

network layer 14. The network layer 14 then juxtaposes a field 13 with the field 7 in order to obtain an acknowledgement datagram, which is routed to the network layer 10. The network layer 10 then transmits the field 7 of the acknowledgement datagram to the transport layer 4 via the interface 11.

Referring to Fig. 1, the network layer 14 routes the datagrams through the physical network 18 to the network layer 16 of the router equipment 15, which reroutes them through the physical network 17 to the network layer 10. When the network layer 16 of the router equipment 15 receives a datagram 24 that it cannot immediately retransmit through the physical network 17 due to congestion, it accumulates the diagram 24 in a second queue 25, which it empties as the physical network 17 becomes available.

A window system of the transport protocol informs the transport layer 4 of the quantity of information it can send to the transport layer 6 prior to receiving the acknowledgement segment. To do this, the transport layer 6 regularly sends segments containing an indicator of the quantity of information it can process without becoming saturated. A simply way to do this is to enter this indicator, for example, into the field 7 of the acknowledgement segments.

Several reasons can cause the transport layer 4 not to receive acknowledgement for segments generated and sent to the transport layer 6. For example, the segments generated and sent to the transport layer or the acknowledgement segments may be lost in the network layers. The transport layer 4 can then re-send unacknowledged segments until it receives acknowledgement for them.

Referring to Fig. 4, the router machine 15 comprises a device for reducing congestion in the network layer 16 when it accumulates, in the queue 20, datagrams 12 to be sent through the network 18. The device comprises means 33 for detecting any datagram received from the network 18 wherein a field 28 of the transport layer 6 contains a received window value VFR, and for entering a sent window value VFE into it based on a fullness level 26 of the queue 20 prior to routing the detected datagram through the network 17 to the transport layer 4. When the fullness level of the first queue 20 exceeds a warning threshold 26, the network layer 16 detects the acknowledgement datagrams 27 coming from the network 18 and processes the content of their field 28 before retransmitting them through the network 17. The

processing of the field 28 is done so that the window value slows the flow of datagrams entering the router 15 addressed to the network 18. The network layer 10 then receives the datagram 27 with a window value in the field 28 that not only takes into account the processing capacity of the transport layer 6 but also takes into account the processing capacity of the network layer 16. The network layer 10 extracts or moves the field 13 of the datagram 27 in order to obtain a segment to be transmitted to the transport layer 4. Based on the window value of the field 28, the transport layer 4 then generates a number of datagrams addressed to the transport layer 6 that is less than or equal to the number of datagrams acceptable by the transport layer 6 of the DTE 2.

Referring to Fig. 6, a method for reducing congestion in a machine comprises four steps. A first step 29 measures the fullness level of the first queue 20 of the datagrams to be sent through the network 18, in order to generate a signal NIV. A second step 30 detects the datagrams coming from the network 18 containing the field 28 with a received window value VFR. A third step 31 processes the value VFR so as to generate a sent window value VFE and to replace the value VFR with the value VFE in the detected datagram, based on the signal NIV. A fourth step 32 retransmits the detected datagram through the destination network 17.

The method, although implemented at the level of the network layer 16 of the router equipment 15, by replacing the window value VFR with the value VFE, modifies a field of the transport layer. It is necessary to comply with the constraints linked to the transport protocol. Step 30 therefore begins by identifying the transport protocol in the field 7 of the datagram received.

For example, in the case of the known transport protocol TCP, the segments are transmitted in byte sequences, each numbered from the first to the last byte in the sequence. Upon receiving the last byte of a sequence, the transport layer 6 of the receiving DTE 2 sends an acknowledgement if this is the first sequence or if it has already sent an acknowledgement for the sequence that immediately preceded it. This acknowledgement generally indicates the number of the first byte of the next sequence waiting to be received. In the same segment that contains the acknowledgement, the receiving DTE 2 sends a window value VFR representing the number of bytes that the sending DTE 1 can send in the sequences to come. The value VFR takes into account any value that may have already been transmitted with a

previous acknowledgement, indicating the number of bytes already received and the number of bytes that are acceptable on the receiving end.

For each connection established between the transport layers 4 and 6 wherein the datagrams of the network layers 10 and 14 pass through the network layer 16 of the router equipment 15, the router equipment 15 detects the upstream transport protocol type. If the transport protocol detected is the TCP type, the router equipment 15 calculates, in parallel with steps 29 through 32, a remaining window value VFER representing the number of bytes that the sending DTE 1 can still transmit at the moment this value is calculated. In order for the remaining window value VFER to represent reality, it is essential to force all the datagrams in the same connection to pass through the router equipment constituted by the machine 15.

The remaining window value VFER is calculated in the following way. Each time the means 33 receive a datagram containing an acknowledgement, the value it indicates is stored in a variable named ACK. A variable ACKp, initialized at zero, contains the value indicated by the previous acknowledgement. A value Diff is calculated by the formula:

$$Diff = ACK - ACKp$$

The value Diff therefore represents a number of bytes sent in a window VFEp previously transmitted to the DTE 1. The value VFER is therefore given by the formula:

$$VFER - VFEp - Diff.$$

The value VFE obtained in step 31 is therefore equal to the larger of two window values VFER and VFI, where VFI is an intermediate window value calculated based on various possible implementations as explained in the description below, the choice of which is left to a network administrator.

$$VFE = \max (VFER, VFI)$$

This makes it possible to ensure that the value VFE is never lower than the window value VFER for which DTE 1 will continue to transmit bytes prior to receiving the new window value VFE.

According to a first possible implementation, in step 29, the signal NIV is set to a binary alarm state when the fullness level exceeds a first threshold. In step 31, when the signal NIV is in an initial state, the value VFI is equal to the value VFR, and it is the transport layer 6 that imposes the window value on the transport layer 4 in

order to regulate its sending of datagrams. Steps 30 and 31 can be short-circuited, i.e., the datagrams with windows can be retransmitted directly from the network 18 to the network 17. When the signal NIV is in the binary alarm state, the value VFI is obtained by taking the lower value of the value VFR and a predetermined offset value VFT based on the capacity of the network 18 to empty the queue 20. This has the effect of momentarily reducing the exchanges in high-speed transport layers 4, 6 without necessarily reducing them in low-speed transport layers 4, 6, whose window values are already lower than the offset value VFT. A variant consists of obtaining the value VFI by multiplying the value VFR by a coefficient of less than one. This has the effect of momentarily reducing the exchanges in all of the transport layers 4, 6 in a proportionally identical way, for both low-speed layers and high-speed layers. When the signal NIV is reset to the initial state, the datagrams with the window value VFR are once again retransmitted normally. The signal NIV is reset to the initial state in step 29 when the fullness level falls below the first threshold or when the fullness level falls below a second threshold lower than the first threshold. The hysteresis thus induced in the limitation of the size of the windows has the effect of preventing instability. The second threshold can be very low, so as to correspond to an empty state of the queue 20.

According to a second possible implementation, in step 29, the signal NIV is the one's complement of a number TAUX obtained by dividing the measured fullness level by the total capacity of the queue 20. Thus, when the queue 20 is empty, the signal NIV is equal to one, and when the queue 20 is full, the signal NIV is equal to zero. In step 31, the value VFI is obtained by multiplying the value VFR by the signal NIV. Thus, when the queue 20 is empty, the value VFI is equal to the value VFR and the datagrams remain unchanged. When the queue 20 is full, the value VFI is null, which means that the transport layer 4 can only retransmit a datagram to the network layer 10 after having received an acknowledgement for a preceding transmitted datagram. Between these two extremes, the size of the windows is progressively reduced, with a value VFI between VFR and zero. In case of a momentary overload of the network 18, the fullness level of the queue 20 has a tendency to stabilize around an intermediate value, which makes it possible to anticipate a subsequent load reduction. It is possible to act on this intermediate value by introducing the number TAUX in polynomial form into the calculation of the signal NIV.